网站新闻全网阅读量统计方法研究

摘　要：网站新闻是网络新闻传播的重要数据源，统计网站新闻在经过网络多次传播后的全网阅读量具有重要意义。然而，目前尚未有成熟的全网阅读量统计方法。本文对网站新闻全网阅读量统计方法展开研究，在分析统计网站新闻全网阅读量面临的各种复杂度的基础上，提出了一个统计算法模型，并分析了该模型的优缺点。

关键词：网站新闻；全网阅读量；统计算法
中图分类号：G203　　**文献标识码：**A
文章编号：1671-0134（2018）08-117-03　　**DOI：**10.19483/j.cnki.11-4653/n.2018.08.048

文 / 陈泰伟　苏国伟　程策

1. 统计网站新闻全网阅读量的意义

在网络媒体、自媒体、移动媒体不断壮大的今天，网站已经在一定程度上成为了传统媒体平台。虽然直接从网站获取新闻的网民在不断减少，但网站新闻一直是各平台网络新闻转发分享的重要数据来源，而且网站新闻在权威性、真实性上相对其他媒体平台具有明显优势。

统计网站新闻传播获得的全网阅读量具有重要意义。从国家层面看，新闻宣传主管机构需要掌握重要政策、权威信息、宣传内容的落地情况；从传媒行业层面看，各新闻媒体单位需要了解自身媒体的影响力，整个行业也需要给出影响力排行；从新闻策划层面看，新时代的策划者已经不能再只凭自身经验和新闻敏感度做出决定，决策必须要有数据参考。以往，各媒体单位更多是依靠自身的网站访问量统计系统获取网站新闻的传播数据，该数据只能代表网站新闻在单个媒体平台的阅读情况，不能反映全网阅读情况。本文提出的全网阅读量，为单个新闻的全网传播效果给出了一个量化指标，进而更能满足各层面对传播效果的统计需求。

另一方面，随着科技的进步，文本相似度计算在信息检索的效率提高方面起到了很大的作用。[1] 再加上目前大数据分析技术的日臻成熟，在对全网进行数据挖掘的基础上，能够通过文本相似度算法跟踪一篇稿件在全网的传播情况，这为统计网站新闻全网阅读量提供了技术可能。

2. 统计网站新闻全网阅读量的复杂度

与统计单个网站的网站新闻阅读量不同，要统计一篇网站新闻稿的全网阅读量，会受到网站新闻稿所在的空间、时间、传播过程以及统计过程等多方面因素的影响，接下来本文从这四个维度加以分析。

2.1 空间复杂度

网站新闻被不断转发后，会出现在网络空间多个位置上。首先，稿件会出现在多个网站上，不同的稿件被转发的网站数量各不相同；其次，稿件可能出现在同一网站的多个位置上，例如在网站首页、网站相关频道首页、网站专题页、网站子栏目页等；再次，稿件还可能在社交网络上有更复杂的存在形式，比如，论坛、贴吧、微博、微信等（关于稿件在社交网络上的阅读数，多可从各平台直接获取，本文统计算法中暂不考虑）。

2.2 时间复杂度

不同时间点稿件的传播情况不同。随着时间变化，稿件逐渐出现在多个网络空间位置上，统计时间点不同，稿件的空间位置数量也不同，统计得到的阅读量也就不同。

不同时间点稿件的热度也不同。诸如热度衰减、再次发酵、旧闻新炒等，导致统计的阅读量也不同。如图 1 是一条真实新闻稿件阅读量随时间变化的曲线图，该图展示了该条稿件从变热到衰减最后到消亡的过程。该新闻稿件从 4 月 30 日凌晨发稿后，在当日 15 点到 19 点较短时间内阅读量达到最大，然后稿件热度衰减，阅读数也随之逐渐下降。在次日的 3 点处于衰减期的该稿件由于某种外界因素被重新激活，稿件阅读量重新上升，然后又开始衰减，最后消亡。

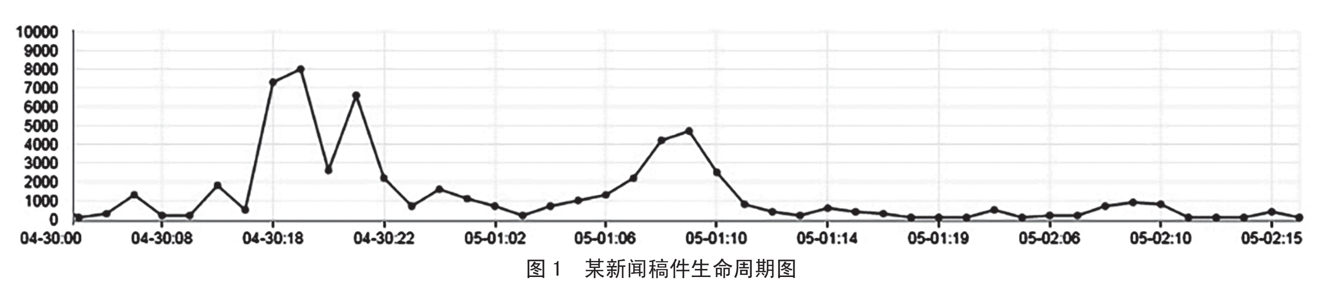


图 1　某新闻稿件生命周期图

2.3 传播复杂度

稿件在传播过程中会面临许多复杂的情况。第一，转载媒体可能会对被转载稿件的标题进行修改，甚至对内容进行增删处理；第二，有的转载媒体并不标注转载来源，造成在溯源统计中稿件传播链的断裂；第三，同一个转载媒体会将同一篇稿件转发到同一网站的多个位置，形成多个传播链分支；第四，稿件被转发后在各个空间位置的停留时长不同，例如稿件在一段时间内出现在某网站的首页大标题上，不久后该稿件从首页大标题上撤下，也就不再具备该网站位置的曝光率和阅读量。

2.4 统计复杂度

在实际统计过程中会面临许多复杂的情况，也会增大全网阅读量统计的难度，大致包含以下一些情况：首先，不是所有网站对自己稿件的阅读量都有统计；其次，即使有的网站对阅读量有统计，各网站的统计方法和标准也不尽相同；再次，一般来讲，大部分网站不会对外公布自己的真实统计数据；还有，就算各网站都公布了自己的统计数据，对全网各统计数据进行收集整理的难度也非常大，几乎很难实现；最后，由于很可能不能及时完整地获取各网站统计数据，各网站统计数据又都在不断随时间变化，使得统计周期长，统计时间点很难把握，最后得到统计结果的时效性和真实性都不大。

3. 统计网站新闻全网阅读量的算法实现

基于以上复杂度分析，要想获取精确的网站新闻全网阅读量几乎是不可能的。但是我们可以通过一定的算法模型估算稿件的阅读量，使计算出的全网阅读量能在数量级上提供参考价值，从而一定程度上解决这个难题。

3.1 统计网站新闻全网阅读量的前置条件

条件一，明确对网站新闻阅读量的定义。本文所指的网站新闻阅读量，指用户通过浏览器打开稿件正文页一次，即算贡献一个阅读量，即页面浏览数（PageView，PV）。

条件二，能够获得被统计稿件在首发网站的阅读量。本算法使用者一般是某个网站媒体，依据本算法计算本网首发稿件的全网阅读量。首发网站通常能够获取自身网站的稿件阅读量，如果不能则可通过在网站后台部署一套访问量统计系统即可实现。本算法将以此作为计算基础，力争提高计算结果的可信度。

条件三，我们假设通过大数据分析，能够获取到稿件被转载的媒体以及该稿件在该转载媒体上所属的栏目。现在大数据技术和网络爬虫技术都趋于成熟，爬取新闻网站的稿件，然后通过相似性算法对比新闻稿件的内容实现对原创新闻稿件的跟踪，从而获取原创稿件被转载的媒体和所属被转载媒体的栏目。

3.2 统计网站新闻全网阅读量的算法描述

为了便于说明，本文以中国军网（以下简称“军网”）的首发新闻稿件为例，对网站新闻全网阅读量统计算法展开分析。

假设现有一篇军网原创稿件被投放到军网的军媒要文要论、军网关注、国内新闻等共 n 个栏目下，军网作为稿件首发网站，一段时间后假如是 t_0 小时该稿件在军网的阅读量表示为 R_0，则 t_0 小时后 R_0 为：

$$R_0 = R_{01} + R_{02} + \ldots + R_{0n} \qquad \text{（公式 1）}$$

其中，R_{01}、R_{02}、R_{0n} 代表不同栏目下的阅读量，这一组阅读量可以从军网自己后台的访问统计系统中获取到。然后，我们计算该稿件在军网单个栏目下的平均阅读量。由于同一稿件不同栏目下的访问量不同，比如出现在网站首页大标题上被点击的概率一定比出现在网站其他栏目点击率高，所以，在计算单个栏目平均阅读量时，不能简单把各栏目的阅读量取平均值作为单个栏目的平均阅读量，我们采用加权平均的算法，这样更符合实际。本文以 r_{0i} 代表军网第 i 个栏目里稿件的阅读量权重值（$0<r_{0i}<1$），则该稿件在军网单个栏目的平均阅读量为：

$$\overline{R_0} = r_{01}R_{01} + r_{02}R_{02} + \ldots + r_{0n}R_{0n} \text{且} \sum_{i=1}^{n} r_{0i} = 1 \qquad \text{（公式 2）}$$

现在，假设有网站 1 此时有多个栏目转载了该篇稿件，并且我们无法获得该稿件在该网站的阅读量。那么，我们可通过引入网站 PR 值比值和网站日均访问量比值的方式估算该稿件在该网站的阅读量 R_1。

引入 PR_1 代表网站 1 在 google 网站的网站 PR 值。网站受欢迎程度越高，该网站越容易被搜索引擎收录，收录数越大网站被访问率就越高。在评价一个网站的受欢迎程度时，我们采用 google 网站为各网站定义的 PR 值来衡量。网站的 PR 值（全称为 PageRank）是 google 搜索排名算法中的一个组成部分，[2]PR 值的级别从 1 到 10 级，10 级为满分。PR 值越高说明该网站越受欢迎。例如，人民网的 PR 为 7、军网为 6、新华网为 9 等。各网站 PR 值可以通过编写程序从 google 网站获得。

引入 A_1 代表网站 1 的日均访问量。不同网站的日均访问量可以通过 Alexa 网站获取。用户通过装有 Alexa 工具栏的浏览器访问某个网站时，Alexa 工具栏就会把访问信息记录并发送到 Alexa 网站，然后 Alexa 网站计算出每个网站的日均访问量。[3] 虽然这个访问量是相对访问量，不是真实访问量，但可以作为计算网站间访问量比值的依据。各网站的 Alexa 日均访问量也可以通过编写程序从网上获得。

PR_1 与 PR_0 的比值为网站 1 相对于军网受欢迎程度的倍数，A_1 与 A_0 的比值为网站 1 相对于军网的日均访问量倍数。

假设 m 为网站 1 在此时转载了该新闻稿件的栏目总数，则此时该稿件在网站 1 的阅读量 R_1 可由以下公式算得：

$$R_1=\frac{PR_1}{PR_0}\cdot\frac{A_1}{A_0}\cdot\overline{R_0}\cdot m \qquad \text{（公式 3）}$$

假设 t_0 这段时间内有 N 个媒体转载了军网的上述原创稿件，则 t_0 这段时间内该稿件的总阅读量 W_{t0} 为：

$$W_{t0}=R_0+R_1+...R_N \text{且} N>0 \qquad \text{（公式 4）}$$

其中：

$$\begin{cases} R_0=R_{01}+R_{02}+...+R_{0n} \\ R_1=\frac{PR_1}{PR_0}\cdot\frac{A_1}{A_0}\cdot\overline{R_0}\cdot m_1 \\ .... \\ R_N=\frac{PR_N}{PR_0}\cdot\frac{A_N}{A_0}\cdot\overline{R_0}\cdot M_N \end{cases} \qquad \text{（公式 5）}$$

随着时间的推移，R_0 会随着时间变化而变大，成为 $R_0(t)$。由于各个转载媒体阅读量 R_i 的统计都是以 R_0 为基础计算出来的，所以，在 R_0 变大后 R_i 也会变大，成为 $R_i(t)$。同时，m_i 代表第 i 个转载媒体下有 m_i 个栏目转发了军网的原创稿，m_i 也会随着时间的变化而变化，成为 $m_i(t)$，变大则说明对应的媒体扩大了转载栏目的个数，变小则说明对应的媒体在部分栏目进行了撤稿。m_i 的变化也会导致 R_i 的变化。

因此，第 i 个转载媒体的阅读量为：

$$R_i(t)=R_i(R_0(t),m_i(t)) \qquad \text{（公式 6）}$$

即：

$$R_i(t)=\frac{PR_i}{PR_0}\cdot\frac{A_i}{A_0}\cdot\overline{R_0(t)}\cdot m_i(t) \qquad \text{（公式 7）}$$

由此，网站新闻的全网阅读量为：

$$W=W(t)=R_0(t)+R_1(t)+...+R_N(t) \qquad \text{（公式 8）}$$

一般来讲，W 不会一直变大，在该网站新闻稿的标题链接逐步从各网站的栏目页退出后，稿件的传播生命周期基本结束（不考虑网民通过搜索引擎再次访问该网页），稿件的全网阅读量就不再增加了。

3.3 算法实际应用

基于以上算法，本文对军网一篇网站新闻稿《55 岁“高龄”被特招入伍，他凭借的是啥？》进行了跟踪统计。该稿件在中国军网发布 t=8 小时后在各个栏目下的阅读量之和 R_0=1440。其中，R_1=940、R_2=415、R_3=85。通过解放军报大数据服务平台分析发现，此时共有 10 个网站对该稿件进行了转载，具体相关参数如下表所示。

转载网站	转载网站 PR 值	转载网站日平均访问量 A_i	对应网站转发栏目数 m_i
中国军网	6	4144000	3
手机新浪网	7	1478000	2
中国网	8	52480000	1
中工网	7	12000	1
新浪新闻	8	378560000	2
国防部网	7	3000	2
荆楚网	8	576000	1
台海网	7	32000	3
第一推	2	28000	2
hao123	6	86831000	3
中宏网	1	256000	1

取军网 3 个原发栏目的权值分别为 0.6、0.3、0.1，代入公式 2 得 $\overline{R_0}$ =697。把上表中的相关数据代入公式 8 中得 W(8)=227,556。则该稿件在发布 8 小时后的全网阅读量为 227,556。

3.4 算法优缺点分析

算法优点：一是本算法充分考虑了网站新闻阅读量统计的时间复杂性、空间复杂性、传播复杂性和统计复杂性，归纳出了可操作的计算全网阅读量的方法；二是本算法以被统计稿件在某个网站的真实阅读量为基础进行估算其他网站的阅读量，使得计算结果更加真实；三是本算法除了对网站本身、网站日均访问量这些因素进行评估，还考虑了首发网站不同栏目对稿件阅读量的影响；四是使用者可以自己对首发网站不同的栏目设置相应的权值，具有一定的灵活性。

算法不足：一是本算法不能准确的算出一篇新闻稿在全网的阅读量，只是在数量级上提供参考；二是对首发网站不同栏目的权值设置没有一个统一的标准，而是由使用者自己设置，既是优点也是缺点。

结语

一篇网站新闻稿的全网阅读量比在单一网站的阅读量能更好地反映其宣传效果，同时也更适合作为影响力评估、新闻策划的参考依据。本文通过仔细考虑影响全网阅读量的各种因素，归纳出了可操作的全网阅读量算法公式，初步实现了在全网范围内跟踪统计一篇稿件的阅读量，为进一步展开网站新闻传播大数据分析打下了基础。媒

参考文献

[1] 王格，吴钊，李向 . 基于全文检索的文本相似度算法应用研究 [J]. 计算机与数字工程，2016，44（4）：567–571.

[2] 焦金涛 . 基于 PageRank 的 Web 挖掘改进算法 [J]. 计算机工程，2009，35（15）：284–285.

[3] 李奉，郑宏 . 从 Alexa 排名的相关参数比较国内 3 种电子期刊网站 [J]. 情报探索，2009（2）：67–70.

（作者单位：中国人民解放军新闻传播中心 ）